Likelihoods for KOPIO?

by Aleksey Sher,
TRIUMF
July 18, 2004

Likelihood method is becoming increasingly common and what is more important proved to be useful. In this note I try to outline what can Likelihoods do for KOPIO and how it should/would/could be done?

Possible use

Alternative to cuts?

Instead of using cuts, one could use Likelihoods, which in theory at least, would be more effective. In our specific case, for **each mode** (signal and background) one must:

- choose the variables (Z_K,$E^*_{\pi^0}$,$M_{\gamma\gamma}$ etc) that are to be included in the general likelihood
- choose what correlations between variables are to be taken into account
- form probability density functions (p.d.f.'s that could be multi dimensional if correlation is to be taken into account) for selected variables from the Fast MC
- calculate general likelihoods to be from each mode for a given event.

Estimation of the effectiveness of the likelihood could be done in a form of a signal vs background plot. Such plot could be obtained by varying the cut on the likelihood and calculating number of surviving signal and background events.

A comment on actual use

Although such exercise would indicate whether our present cuts are close to optimum and would tell how much could be gained from the likelihood use, there is one complex question to be looked into.

In the real data analysis we would have to select **clear samples of the backgrounds** in order to form the p.d.f.'s and that could be a difficult task.

Optimization

Presently, most of the background studies are carried out with the help of the Fast MC, and are evaluated on the basis of the expected number of signal events and S/B ratio. The number of signal and background events are estimated using a pre-defined set of cuts. Presently we have at least seven such sets (MZ, AK basic to tightest). Moreover, these sets were not changed for quite long time, and re-optimization would be a reasonable thing to do.

As an alternative one could use the likelihood. Result of likelihoods can be given in the form of the number of signal vs the number of background events. It would help to reflect the effects of changes in geometry of the detector on the acceptance of signal and background. Changes in resolution would be also taken into account, since they would be reflected in the p.d.f.'s.

Downsides are the following: to use the likelihood one need large sample of simulated events to derive the PDFs from. In addition, one should use different samples for forming PDFs and testing the likelihood method (or some division of the sample that is available).

Possible requirements for implementation:

- Universality - structure that could accommodate output from the Fast MC, GEANT, and possibly data!?
- Optimization - fast running (p.d.f.'s are regenerated only by request of the user, etc.)
- Flexibility - can easily add variables and change correlation structure

It is possible to add an option of running a likelihood analysis, not a likelihood cut.

Current structure

- INITIALIZATION:
  - Input parameters are: variables, correlations, mode
  - reconstructed variables are written out into a file from the modified **anal.x** routine (analysis of the FastMC).
- STAND ALONE routine derives PDFs from the file. Weights are taken into account
- LIKELIHOOD CALCULATION:
  - Likelihood to be from *ith* mode is calculated for each event
  - Weights are taken into account, like in the current analysis scheme
- RESULT:
  - Number of events from each mode is calculated as a function of the relative likelihood cut (relative against *ith* background).

Defining Variables and relations

Here is a list of variables that are used to define the general Likelihood variables and their relations:

NVAR (int) - number of variables to be used
NDimMax (int) - Maximum number of dimensions to describe the variables relations
N1Dim - Number of 1-Dimensional variables
N2Dim - Number of 2-Dimensional variable pairs
Dim1 (n) - Array that stores 1-dimensional variables indexes
Dim2 (2,m) - Array that stores 2-dimensional variables, indexes
*VarName (NVAR,char*15) - Name of the variables (not essential)*

From this variables, Likelihood program defines p.d.f.'s.

Current problems

- Derived PDFs are not smooth, because of the event weight (caused by VETOes inefficiencies, which can range by orders of magnitude). Refer to Fig 1 for details.
- Work on binning algorithm is in progress. Bin with zero events in them cause problems or unrealistically good results. Have some ideas.
- Currently got to have two MC samples for each mode (first one is used to form the PDFs, second is used for testing the likelihood). Possibly devision of the sample must be an option.

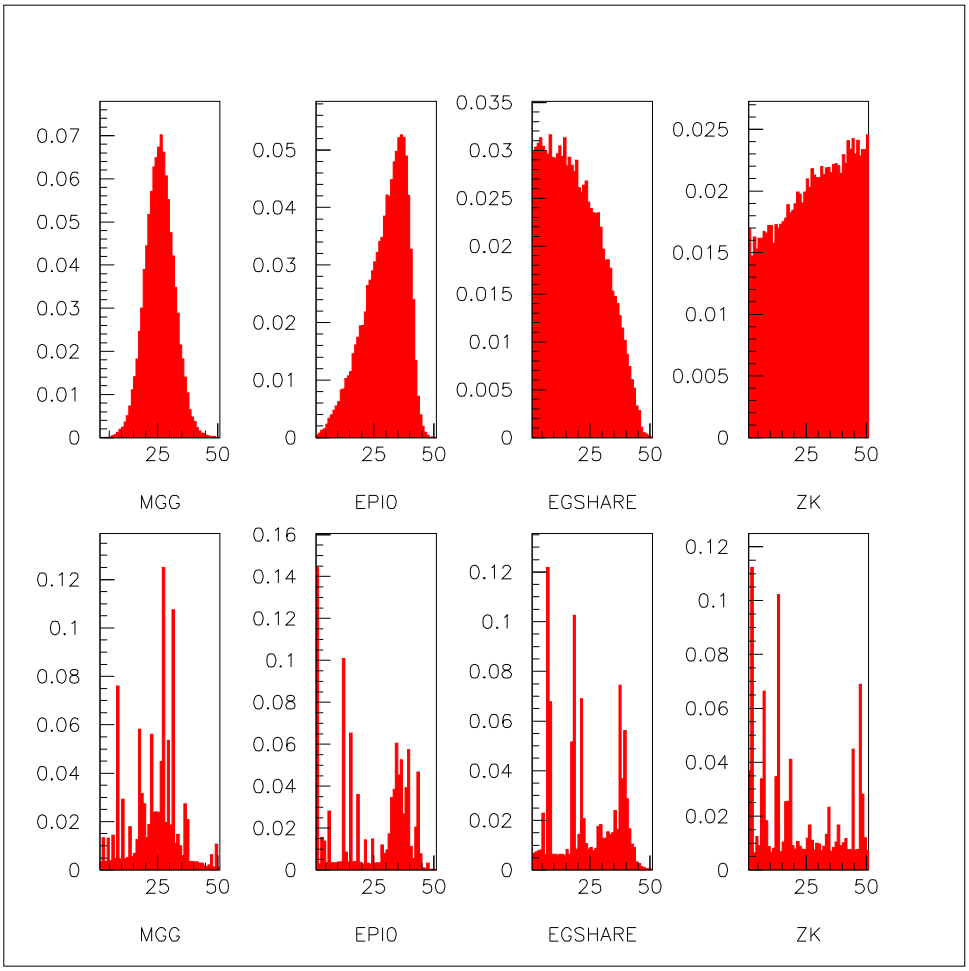

Figure 1: .Signal (top) and kp2 (bottom) background PDFs, derived from the FastMC.